Exploring Massive, Genome Scale Datasets with the GenometriCorr Package

Alexander Favorov[1,2,3☯*], **Loris Mularoni**[1☯¤a], **Leslie M. Cope**[1], **Yulia Medvedeva**[2,3¤b], **Andrey A. Mironov**[4,5], **Vsevolod J. Makeev**[2,3], **Sarah J. Wheelan**[1*]

1 Department of Oncology, Division of Biostatistics and Bioinformatics, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America, **2** Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia, **3** Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, Russia, **4** Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia, **5** Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

Abstract

We have created a statistically grounded tool for determining the correlation of genomewide data with other datasets or known biological features, intended to guide biological exploration of high-dimensional datasets, rather than providing immediate answers. The software enables several biologically motivated approaches to these data and here we describe the rationale and implementation for each approach. Our models and statistics are implemented in an R package that efficiently calculates the spatial correlation between two sets of genomic intervals (data and/or annotated features), for use as a metric of functional interaction. The software handles any type of pointwise or interval data and instead of running analyses with predefined metrics, it computes the significance and direction of several types of spatial association; this is intended to suggest potentially relevant relationships between the datasets.

Availability and implementation: The package, GenometriCorr, can be freely downloaded at http://genometricorr.sourceforge.net/. Installation guidelines and examples are available from the sourceforge repository. The package is pending submission to Bioconductor.

Citation: Favorov A, Mularoni L, Cope LM, Medvedeva Y, Mironov AA, et al. (2012) Exploring Massive, Genome Scale Datasets with the GenometriCorr Package. PLoS Comput Biol 8(5): e1002529. doi:10.1371/journal.pcbi.1002529

Editor: Hilmar Lapp, National Evolutionary Synthesis Center, United States of America

Received August 24, 2011; **Accepted** April 8, 2012; **Published** May 31, 2012

Copyright: © 2012 Favorov et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The work was supported by Russian Foundation for Basic Research (grant 11-04-02016-a to AF, 09-04-92742 to AM), by the Presidium of Russian Academy of Science Program in Molecular and Cell Biology (VJM, AM) and in Basic Science for Medicine (AM), by "Russian Ministry of Science and Education State Contract (07.514.11.4005)" (VJM), by National Center for Research Resources grant (UL2RR025005 and supplement, UL1RR025005) (SJW), by the Johns Hopkins University Framework for the Future (SJW), and by the Commonwealth Foundation (SJW). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: favorov@sensi.org (AF); swheelan@jhmi.edu (SJW)

¤a Current address: Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain.
¤b Current address: Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

☯ These authors contributed equally to this work.

This is a *PLoS Computational Biology* Software Article

Introduction

Manual exploration of high-dimensional whole-genome datasets is possible, to a limited extent, with newer, high-capacity genome browsers. While biologists can browse their data and can often suggest relevant hypotheses for statistical testing, fully informed and thorough data exploration is impossible to do by eye.

A common theme in biological experiments is that the nucleotide-level proximity of a set of genomic regions (points or intervals) to genome annotations or to other experimentally derived data sets (such as coverage peaks, mutation locations, and breakpoints) is a useful proxy for a functionally relevant or otherwise interesting interaction. For example, the well established overlap of CpG islands with the promoter regions of genes [1] is critically related to the gene-silencing mechanism of DNA hypermethylation.

While using spatial proximity to infer functional relationships is a valid approach in many cases, this is not necessary for functional interaction, as chromatin is flexible and many activating and repressive marks act at a distance [2], so ideally any software that attempts to automatically uncover important relationships should be sensitive to these interactions as well. Others have given thought to examining some of the interactions that we will discuss, (for a review see Bickel et al, 2009 [3]); however, the only software for performing such analyses focuses on overlapping features [4].

Here we present a method for identifying whether two sets of intervals are spatially correlated across a genome, detected as a deviation from a nonuniform distribution of one set of intervals with respect to the other. This is not a trivial task, computationally or conceptually. Many different spatial rearrangements are possible, each with different biological implications and each requiring specialized statistical analyses. The software performs all analyses on each input, so that a variety of biologically significant relationships are queried. This includes looking for proximity, looking for uniform spacing, looking for increased or decreased overlaps of intervals or points, and presenting the data in a way that a biologist can understand. Results from each test are provided for each chromosome from the dataset and for the entirety of the dataset, which in most cases is the entire genome. As we have found that asking "is A related to B" is completely different from asking "is B related to A," we encourage users to not only perform all comparisons but to perform them in both orientations.

Design and Implementation

Overview and general considerations

GenometriCorr is written in R (using S4 classes) and makes use of the Bioconductor [5] packages IRanges and GRanges to create sets of intervals that are then compared. The input data can be in a variety of commonly used biological data formats, and the core functions work from a configuration file that sets parameters in a straightforward, easy to edit format that can be archived to ensure reproducibility. We provide a Tk interface so that non-programmers may access the functions via straightforward menus, and we also provide a Galaxy [6] plugin that runs the analysis in an environment widely used by biologists, in which the results may be explored more thoroughly. Finally, we provide two auxiliary methods for output, so that graphical results can be obtained in addition to the statistics produced by the main function. The configuration and the result, GenometriCorrConfig and GenometriCorrResult are designed such that once a configuration file has been read, the software proceeds with a simple run.config call.

The main function, GenometriCorrelation, implements various statistical approaches to assess whether the positions of two sets of intervals are associated in genomic space. As stated above, almost all of the tests are asymmetrical, in that one of the two interval sets is considered to be a reference, fixed in the genome, while the other set, the query, is evaluated statistically with respect to the fixed reference. The results can be very different if the reference and query sets are swapped, as shown below. In essence, each of the tests is designed to evaluate whether the spatial distribution of the query intervals is independent of the positions of the reference intervals, and each test is sensitive to a different aspect of known biological relationships.

Two types of graphical output are produced. Calling the graphics.plot() function produces a straightforward statistical summary and ECDF plot for the relative and absolute distances for each chromosome and/or the entire genome. Summary results for all chromosomes together are displayed in a window and results for the individual chromosomes are written to a PDF. The visualize() function produces a more elaborate and biologist-friendly color-coded density plot, intended to represent areas of high and low absolute and relative distance correlation; again, summary results appear in a window and chromosome-by-chromosome results are written to a PDF. The two types of output are shown in Figure 1, along with detailed descriptions of the features of each. These data are Hermes transposon insertions in the yeast genome; this transposon generally inserts into nucleosome free regions [7].

An important consideration here is that two sets of genomic features may only be correlated in one direction. As an example, we found all NF-kappa-B (NFKB1) sites using a simple exact string search of the human genome and correlated their positions to the positions of all RefSeq gene [8] start sites. NF-kappa-B is a family of transcription factors critical in many processes, including immunity, inflammation, and cell growth [9].

As there are nearly five times as many transcription start sites as potential NF-kappa-B sites, most transcription start sites will not be near a NF-kappa-B site even if they are perfectly correlated, while the NF-kappa-B sites will nearly all be close to transcription start sites. Figure 2a depicts the excellent correlation between human NF-kappa-B sites to transcription start sites (same distribution, perfect absolute distance correlation), and Figure 2b depicts the poor (and not statistically significant) correlation of transcription start sites to NF-kappa-B sites (absolute distance indicates a separation, K-S not significant). As this level of asymmetry is common, if not expected, in biological datasets, we recommend performing all comparisons in both directions, using each dataset as the fixed set in turn. While the relevant comparison is not known a priori, the results of the two comparisons will clearly indicate whether the relationship is asymmetric.

Many of the tests we used are performed on pointwise representations of the intervals rather than on the intervals themselves. When the input includes intervals, the midpoints of these intervals are used for those analyses. Very large intervals may relate to genomic features in different ways, depending on whether we examine their start points, stop points, both boundaries, or just a point in the middle. Rather than trying to address this ambiguity or to randomly guess at what the user hopes to do, we expect the user to specify the points when the exact point is important, and we use the midpoint when the user inputs an interval.

We have developed and tested four useful and relevant metrics, which will be discussed below: the relative distance test, the absolute distance test, the Jaccard test, and the projection test, intended to measure a variety of biologically relevant correlations. In Figure 3 we summarize the metrics and their uses, and in subsequent figures we demonstrate the utility of each type of test, using both published and novel observations. Each figure shows both a standard histogram representation of the relationships between the query and reference sets, in addition to the results and p-values generated by our software. As a strong correlation between the query and reference may involve just a subset of a very large number of points, a histogram of the absolute or relative distances can be uninformative, while the tests performed by the software are sensitive to true correlations within large and overall not strongly correlated datasets. All p-values cited are computed by the GenometriCorr functions. For each test we describe a published dataset for which the test is particularly useful. Table 1 and table 2 summarize the results.

Relative distance test

The relative distance test measures whether two sets of positions are closer together or further apart than expected, where the exact distances are not as important as the relative relationship. For example, a recent publication [10] reported that transposable elements found in genes tend not to be located near splice sites. We tested this association with the GenometriCorr software and found that, first, the two entities do not overlap (the Jaccard and projection tests, summarized in figure 3 and described in detail later, are both significant and in the lower tail) and that both the relative and absolute distance tests show a correlation. Upon closer examination, the transposable element and splice site positions are actually negatively correlated in terms of relative distance; that is, the two types of genomic features tend not to co-occur but are consistently spaced apart, though not by a uniform distance (Figure 4). Results are shown for Alu elements but the relationship holds true for both L1 and Alu elements, in agreement with the reported trends.

The idea behind the relative distance test is that if the query locations are independent of the references, then the relative position of each query point, with respect to the adjacent reference points, will have a uniform distribution. Thus, the null distribution for relative distance test as formulated here is simply a straight line at $y = 0.5$. For this test, intervals are represented as points, located

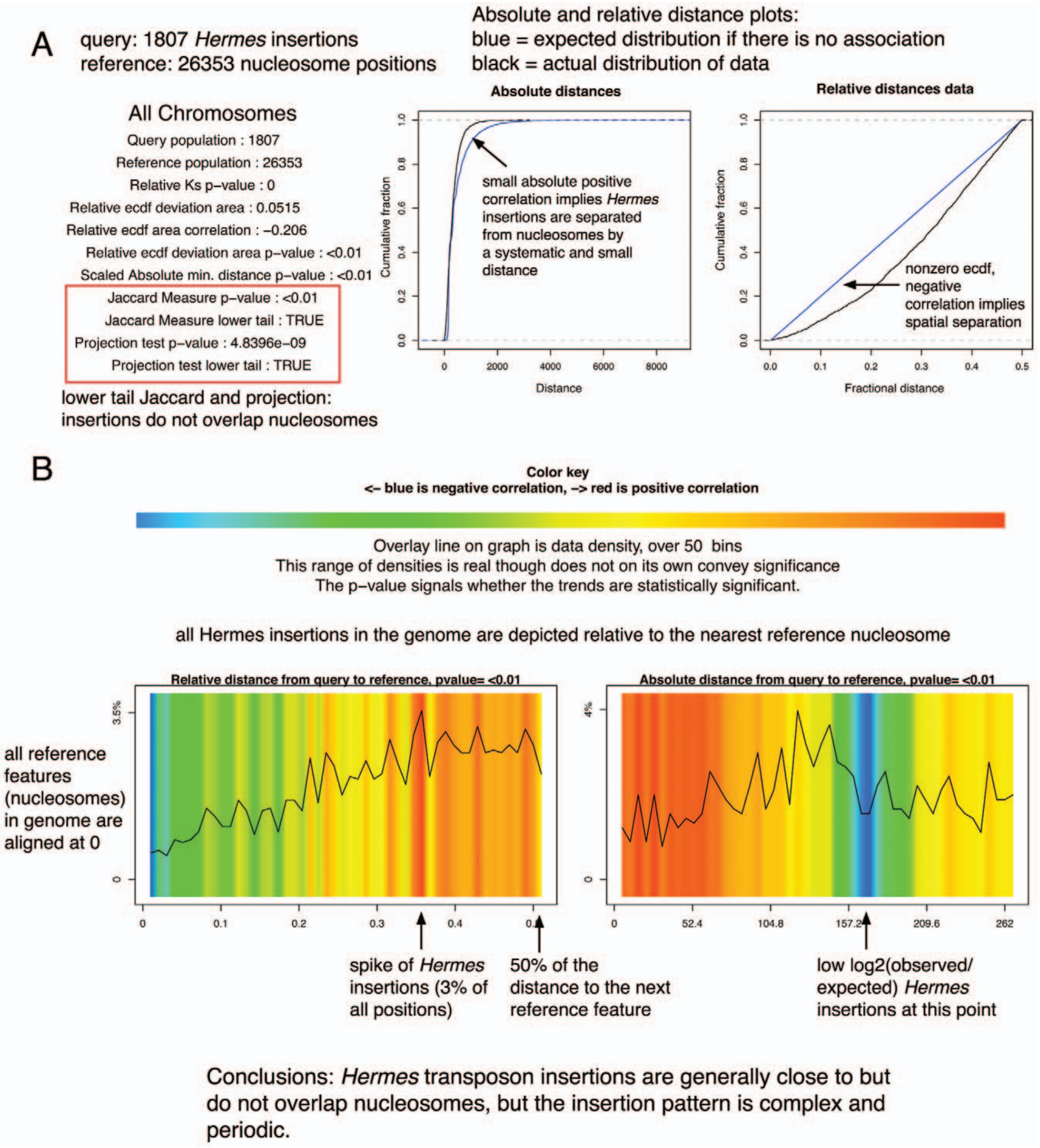


Figure 1. Two types of graphic output are available. (A) A statistical summary and ECDF plots. (B) A graphical interpretation of the spatial relationships. The query features are depicted along the plot according to their distance to a reference feature; the colors indicate deviation from the expected distribution while the overlay line indicates the density of the data at each absolute or relative distance. The data density mirrors but is independent from the log-odds colors; at small distances in the absolute distance plot the data density is higher than expected but this represents a very small percentage of the total query points.

doi:10.1371/journal.pcbi.1002529.g001

at the midpoint of the interval, so that the size of intervals and overlap between query and reference are not included in the analysis.

For each query point, q_i, we identify the flanking reference points, r_k and r_{k+1}, and calculate the relative distance $\delta_i = (q_i, - r_k)/(r_{k+1}, - r_k)$, comparing this to a uniform null distribution. More formally, the "relative distance" δ_i for the i-th query point is:

$$\delta_i = \frac{min(|q_i - r_k|, |r_{k+1} - q_i|)}{|r_{k+1} - r_k|};\quad k = arg \min_{q_i > r_k} (q_i - r_k),$$

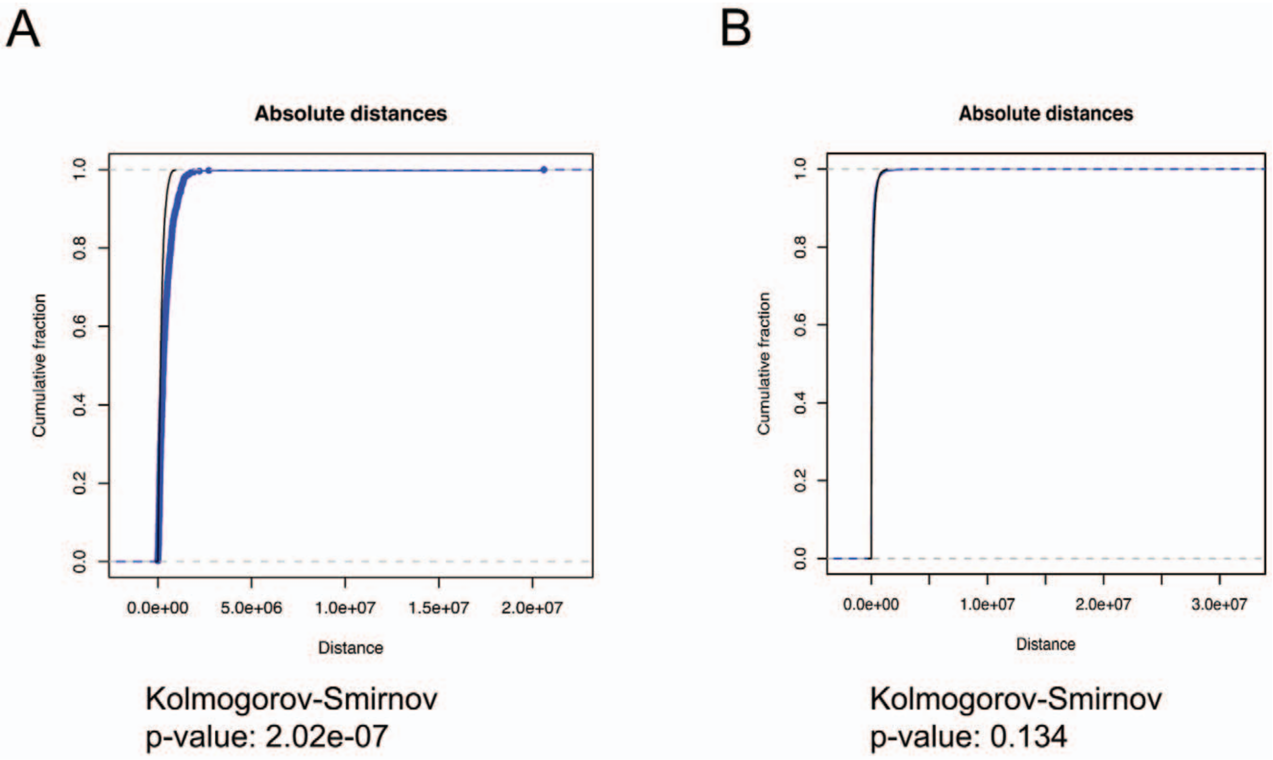


Figure 2. NFkappaB sites vs human RefSeq promoter start sites. Query and reference colors as in Figure 1. (A) NFkappaB as the query gives a significant Kolmogorov-Smirnov association and anticorrelation that is visible from the graph, in absolute distances. (B) Correlation in the reverse direction suggests no significant relationship between the two classes of sites.
doi:10.1371/journal.pcbi.1002529.g002

and under the null, the δ_i's would be distributed uniformly in [0, 1/2].

Two different tests are available to evaluate the uniformity of the distribution the δ_i's. The first and simplest is the standard Kolmogorov-Smirnov test, assessing the maximum difference between CDFs.

A permutation-based test is carried out as well, to more comprehensively compare the two cumulative distribution functions using the area of the region in which they differ as the test statistic. Here *ECDF* (*Empirical Distribution Cumulative Function*) designates the observed distribution of relative distances δ_i, while $ECDF_{ideal}$ describes the expected distribution under the uniform, null distribution, which is a straight line. The area between the *ECDF* and $ECDF_{ideal}$ is then calculated as:

$$S = \int_0^{1/2} |ECDF(\delta) - ECDF_{ideal}(\delta)| d\delta$$

and a *p-value* for S obtained by permutation, in which we randomly draw Q (number of query points) values from the ideal uniform distribution of δ_i and calculate the area S_p.

The integral of the difference between the *ECDF* and the $ECDF_{ideal}$ also permits us to derive a natural measure of association for the two sets of intervals. The sign of the integral indicates the direction of the correlation, so the positive sign indicates that δ_i's tend to be low and thus the query intervals are attracted to reference intervals while, vice versa, the negative sign suggests that query intervals avoid reference intervals. With appropriate rescaling, we define a correlation-like measure:

$$Corr_{ecdf} = \frac{\int_0^{1/2} (ECDF(\delta) - ECDF_{ideal}(\delta)) d\delta}{\int_0^{1/2} ECDF_{ideal}(\delta) d\delta}$$

to express this. The $Corr_{ecdf}$ equals zero for independent query and reference; it equals 1 if each query point coincides with a query point and, finally, it equals -1 if each query point falls in the middle of the reference gap.

Absolute distance test

In some cases, particularly in small genomes in which reference points are closely spaced, the relative distance test produces misleading results. For example, if the promoters in a genome are generally found roughly 100–1500 bases apart (for example, yeast), an element that is positioned consistently 500 bp from promoters will look uncorrelated with promoters in relative terms, as it will sometimes be extremely close to a promoter that is not the one it is functionally related to, thereby diluting the distribution of query-reference distances with many incorrect data points. In these situations the absolute distance test is useful. We created a toy dataset for this analysis, first taking the positions of the start points of all human promoters (31083 sites), creating a new set of small intervals placed randomly from 10–10000 base pairs from each promoter start, and adding an additional 3000 small intervals randomly placed between 75 and 100 bp from a promoter. We then compared these intervals to the actual promoter intervals in the human genome, and the software uncovered the signal of the fixed distance points within the shifted points, whereas a simple histogram approach fails (Figure 5). The null distribution for the

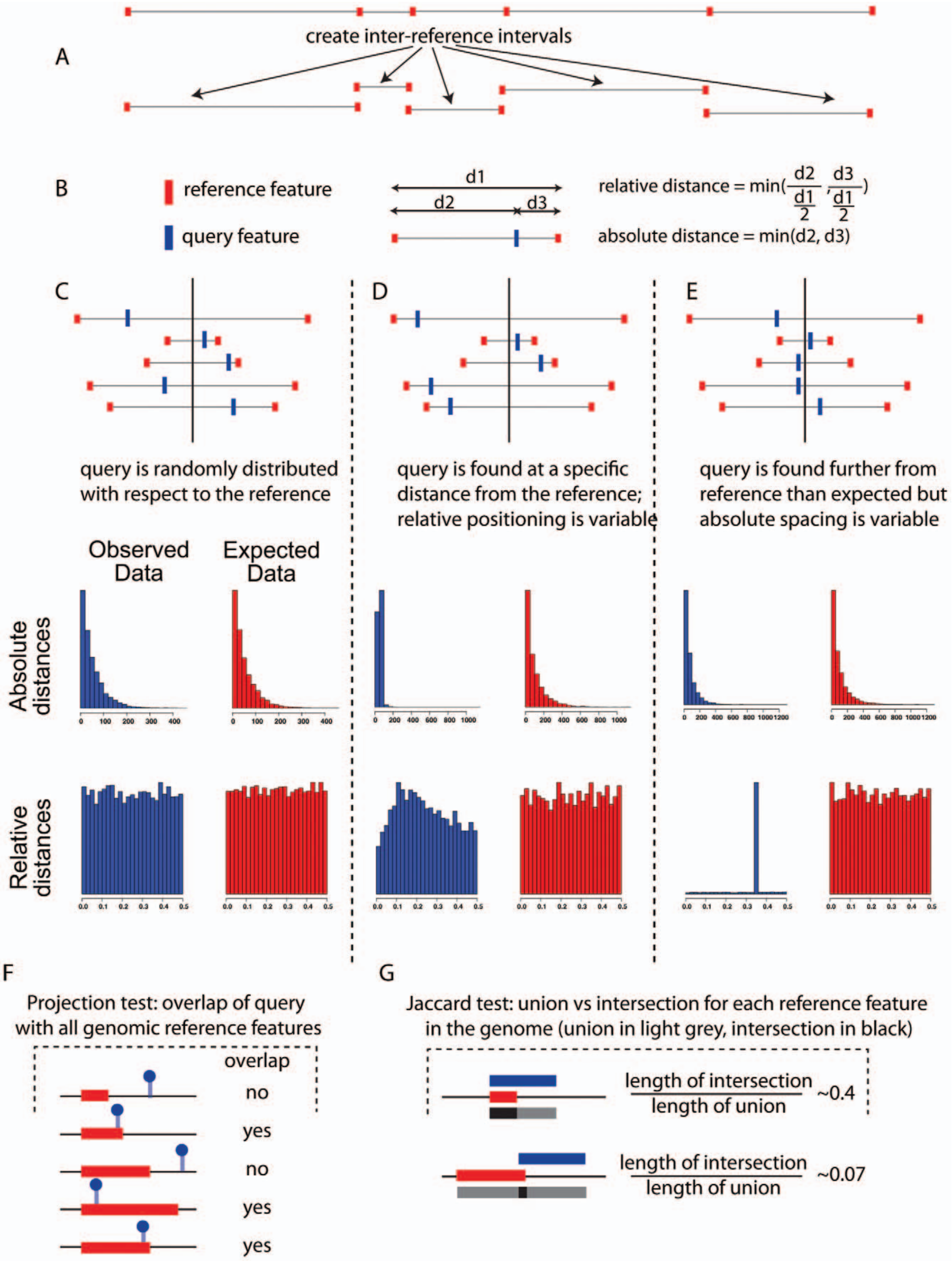

A
create inter-reference intervals
B
reference feature
query feature
d1
d2
d3
relative distance = min(d2/(d1/2), d3/(d1/2))
absolute distance = min(d2, d3)
C
query is randomly distributed with respect to the reference
D
query is found at a specific distance from the reference; relative positioning is variable
E
query is found further from reference than expected but absolute spacing is variable
Observed Data
Expected Data
Absolute distances
Relative distances
0 100 200 300 400
0 200 400 600 800 1000
0 200 400 600 800 1000 1200
0.0 0.1 0.2 0.3 0.4 0.5
F
Projection test: overlap of query with all genomic reference features
overlap
no
yes
no
yes
yes
G
Jaccard test: union vs intersection for each reference feature in the genome (union in light grey, intersection in black)
length of intersection / length of union ~0.4
length of intersection / length of union ~0.07

Figure 3. A schematic of the various tests implemented in the software package, showing when certain tests are most useful. (A) depicts the intervals created in silico and (B) shows how the query distances are evaluated within the intervals. (C) depicts a random distribution of

query versus reference intervals; here the observed and expected distances for both the absolute and relative tests are the same. In (D) we show a relationship best uncovered by the absolute distance test; useful especially for small genomes, this test determines whether the query and reference are often separated by a fixed distance. In (E), the query points are consistently far away from the reference points, so the relative distance test will be significant, while the absolute distances are not significant in this case. Interestingly, the query intervals are variable enough in size that even though the query and reference points are usually separated, the absolute distances between them vary widely in size, including some fairly small distances. (F) demonstrates the projection test, which evaluates whether pointwise data falls consistently inside or outside of a set of intervals. Finally, in (G) we see the Jaccard test, which looks for significant overlaps between datasets by evaluating the ratio of the intersection of the datasets (dark grey) to the union of the datasets (light grey). Perfect correlation will give a ratio of 1, and perfect anticorrelation will result in a ratio of zero.
doi:10.1371/journal.pcbi.1002529.g003

absolute distance test depends on the data. If the inter-reference intervals are somewhat randomly distributed, then the null distribution will be exponential, but if the inter-reference intervals are constrained somehow, the null distribution will have a very different shape.

As in the relative distance test, intervals are represented by their midpoints.

The statistic used is very simple and intuitive. We suppose there are Q query points and R reference points on a chromosome, and for each query point i, find the distance to the closest reference point, scaled by the expected inter-reference gap for the chromosome

$$d_i = \min_k (|q_i - r_k|) \frac{R}{\text{Length of chromosome}}.$$

The final test statistic is the mean value of d_i

$$D = \frac{1}{Q} \sum_{i=1..Q} d_i,$$

which characterizes the spatial association between the query and the reference points. The lower it is, the closer together they tend to be.

The *p-value* is obtained by permutation test. At each iteration, we draw Q simulated query points uniformly distributed along the chromosome and calculate the permuted statistic D_p. The *p-value* is the proportion of permuted statistics that exceed the observed D. As implemented in the package, the test is two-sided and returns both the *p-value* and the direction of the association.

Projection test

Another test included in the software is the projection test. As seen in figure 3, this tests whether pointwise data overlap interval data in a significant way. To confirm the biological relevance of this test we examined data from the Roadmap Epigenomics Project [11]. Using the RNAseq data and histones H3K27me3 and H3K4me3 ChIP data from UCSF-UBC (GEO accessions GSM484408 (RNAseq), GSM428295 (H3K27me3), and GSM410808/GSM432392, (replicates for H3K4me3)), we used the projection test to examine the relationship between the two histone marks and the promoters of the most highly expressed genes. The software was able to determine that the H3K4me3 marks significantly overlap the gene positions (Figure 6A) and the H3K27me3 marks are significantly underrepresented near active genes (Figure 6B). Note that in both cases the projection test is highly significant, but in opposite directions; for the H3K4me3 data the projection test is in the lower tail, indicating significant overlap, while the opposite is true for the H3K4me3 data, indicating very little overlap with promoters of active genes.

The projection test is the methodologically simpler of the two overlap tests in the package; the other, the Jaccard measure, is discussed below. For this test, query intervals are still represented as midpoints, but the reference should be a set of intervals. If the query points are independent of the reference, then the probability that any one query point is contained in a reference interval is the proportion of the chromosome covered by reference intervals:

Table 1. Summary of all correlations performed in these experiments.

	Direction	Relative Kolmogorov-Smirnov p–value	Relative ECDF area correlation	Relative ECDF deviation area p–value		
					C	F
Human transcription start sites (T) versus NF-kappa-B sites (N)	T→N	2e−07	0.015	<0.001	√	
	N→T	0.13	0.012	0.072	√	
L1 elements(T) versus Splice Sites (S)	T→S	~0*	−0.03	<0.001		√
	S→T	~0	−0.16	<0.001		√
Promoter sites (P) versus Promoter plus spikein (S)	P→S	~0	0.25	<0.001	√	
	S→P	~0	0.25	<0.001	√	
H3K4me3 histones (H) versus Promoters of actively transcribed genes (P)	H→P	~0	0.22	<0.001	√	
	P→H	~0	0.02	<0.001	√	
CpG Islands (I) versus Coding sequences (C)	I→C	~0	−0.195	<0.001		√
	C→I	~0	−0.012	<0.001		√

*~0 means that the default R precision for KS test p-value is not enough to distinguish the value from 0.
doi:10.1371/journal.pcbi.1002529.t001

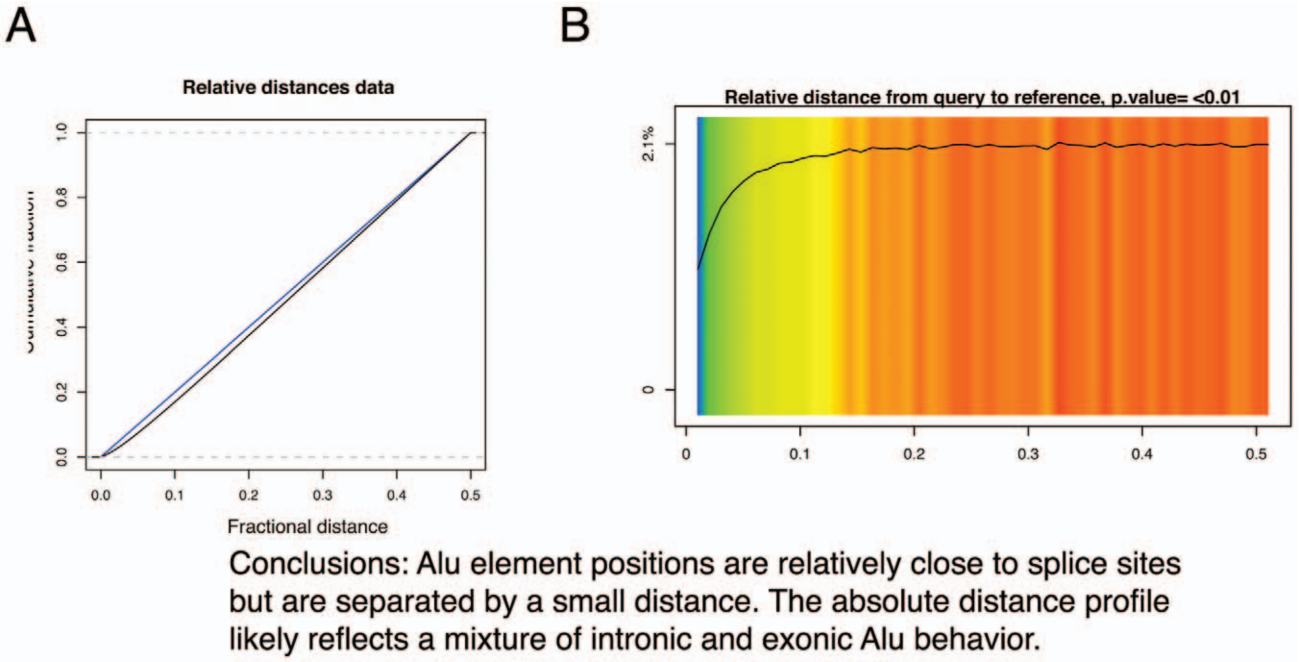
Table 2. Continued summary of correlations performed in the experiments described.

	Direction	**Scaled Absolute minimal distance** ***p–value***			**Jaccard Measure** ***p–value***			**Projection test** ***p-value***		
			C	**F**		**O**	**N**		**O**	**N**
Human transcription start sites (T) versus NF-kappa-B sites (N)	T→N	<0.001	√		<0.001		√	0.0091		√
	N→T	<0.001	√		<0.001		√	~0*		√
L1 elements(T) versus Splice Sites (S)	T→S	<0.001		√	<0.001		√	~0		√
	S→T	<0.001		√	<0.001		√	~0		√
Promoter sites (P) versus Promoter plus spikein (S)	P→S	<0.001	√		<0.001		√	~0		√
	S→P	<0.001	√		<0.001		√	~0		√
H3K4me3 histones (H) versus Promoters of actively transcribed genes (P)	H→P	<0.001	√		<0.001	√		~0	√	
	P→H	<0.001	√		<0.001	√		~0	√	
CpG Islands (I) versus Coding sequences (C)	I→C	<0.001		√	<0.001		√	~0		√
	C→I	<0.001		√	<0.001		√	~0		√

*~0 means that the default R precision for the binomial test is not enough to distinguish the value from 0.

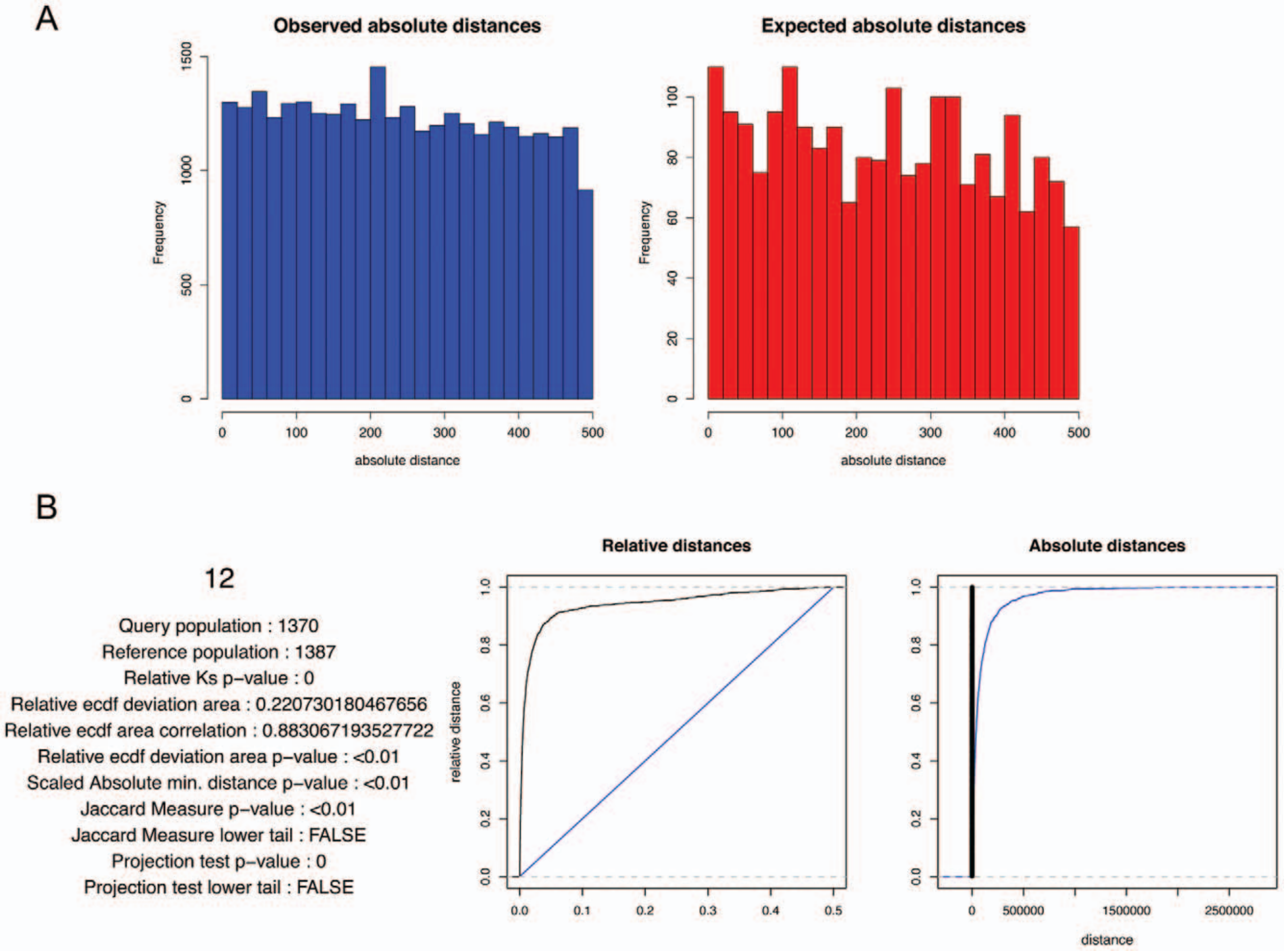
P-values are shown for all tests in both directions (using one dataset as the query and the other as the reference, then reversing). For each test, we have indicated whether the relationship between the datasets is positive or negative; for the relative and absolute distance tests this is written as ''close (C)'' vs ''far (F)'' and for the Jaccard and projection tests it is written as ''overlapping (O)'' vs ''nonoverlapping (N).''

doi:10.1371/journal.pcbi.1002529.t002



Conclusions: Alu element positions are relatively close to splice sites but are separated by a small distance. The absolute distance profile likely reflects a mixture of intronic and exonic Alu behavior.

Figure 4. Alu elements vs splice sites in the graphics.plot() output (A) and in the visualize() output (B). Alu elements are consistently located at a variable but always nonzero distance from splice sites. Query and reference colors as in Figure 1.
doi:10.1371/journal.pcbi.1002529.g004



```
12

Query population : 1370
Reference population : 1387
Relative Ks p−value : 0
Relative ecdf deviation area : 0.220730180467656
Relative ecdf area correlation : 0.883067193527722
Relative ecdf deviation area p−value : <0.01
Scaled Absolute min. distance p−value : <0.01
Jaccard Measure p−value : <0.01
Jaccard Measure lower tail : FALSE
Projection test p−value : 0
Projection test lower tail : FALSE
```

Figure 5. A toy example of absolute distance correlation. (A) Histograms of the observed and expected ranges of minimum distances between the reference and query. (B) GenometriCorr's simple plot for the same data. Query and reference colors as in Figure 1.
doi:10.1371/journal.pcbi.1002529.g005

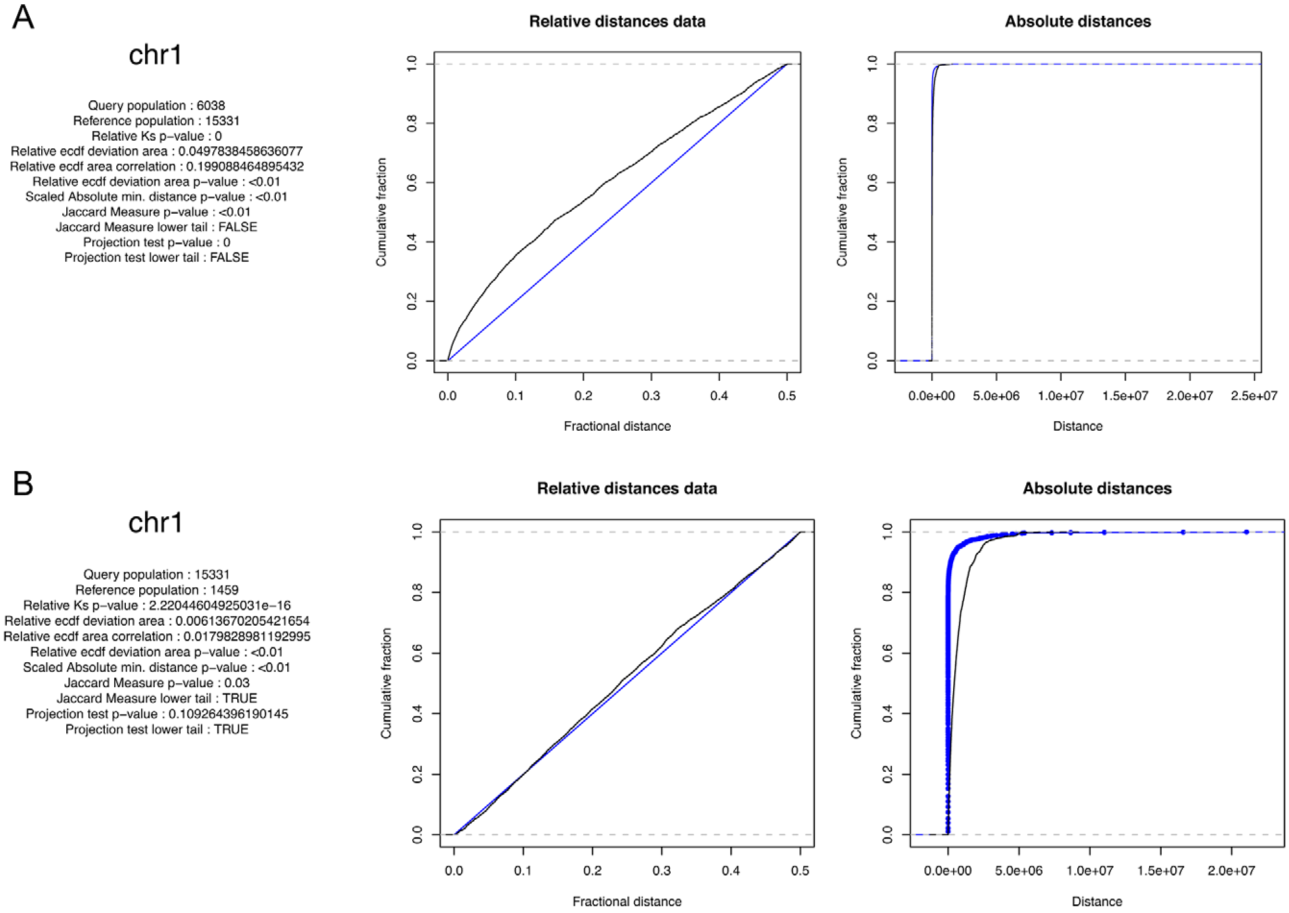


Figure 6. Promoter positions from highly expressed genes (as given from mRNAseq data) and histone ChIP data recently available from the Roadmap Epigenomics Project [8]. (A) H3K4me3 versus highly expressed genes. (B) H3K27me3 versus highly expressed genes. Query and reference colors as in Figure 1.
doi:10.1371/journal.pcbi.1002529.g006

$$p = \frac{\text{coverage of the reference}}{\text{the chromosome length}}.$$

The total number of query points Q_R contained in reference intervals has a binomial distribution,

$$p(Q_R|\text{independence}) = \binom{Q}{Q_R} p^{Q_R}(1-p)^{Q-Q_R}.$$

A standard two-sided binomial test is used to evaluate statistical significance. The test is unlikely to be informative if the genomic coverage of the reference intervals is very close to 0 or 1. ; here the p-values will be extremely high.

Jaccard test

The Jaccard test measures overlaps between two interval sets by measuring the extent of intersection between two interval sets, divided by the length of their union. Thus, two datasets that overlap perfectly have a union that is equal to their intersection, and the ratio is 1. This proves to be a very useful measure for biological data, as demonstrated in Figure 7, in which CpG islands [12] are compared with coding sequences in the human genome. Comparing the CpG islands with the coding sequences we see that the two interval sets overlap much less than expected, given the amount of the genome that each occupies, and this anti-correlation is statistically significant. This is expected, as CpG islands generally occur in promoters and other non-genic regions.

The Jaccard statistic is calculated on intervals rather than points, and is the ratio of the number of nucleotides in the intersection of the reference and query, and the total number of nucleotides spanned by the reference and query together.

More formally:

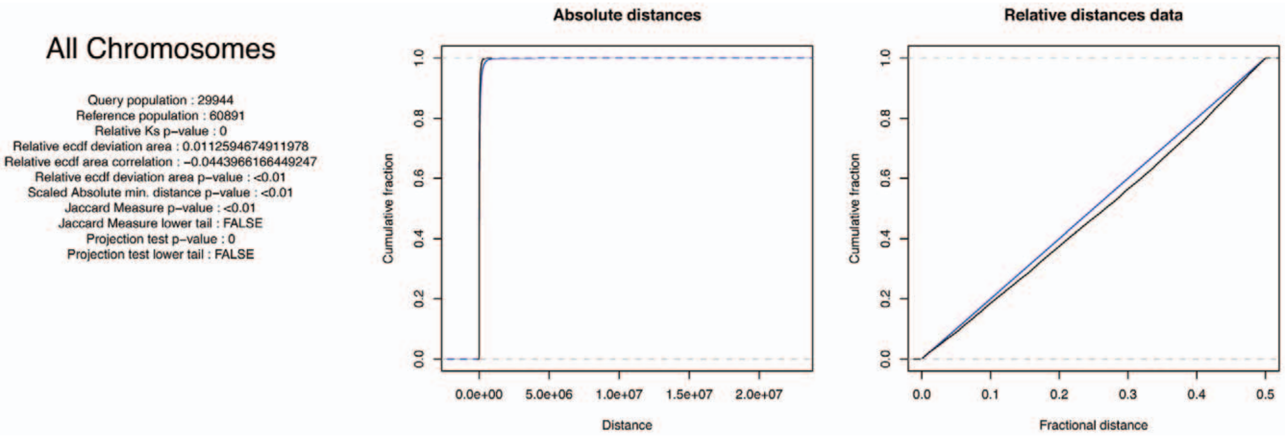
The Jaccard statistic, J, evaluates interval sets $\{Q\}$ and $\{R\}$ that are sets of chromosome positions that are covered by query and reference intervals, respectively.

$J(\{Q\},\{R\}) = \frac{|\{Q\}\cap\{R\}|}{|\{Q\}\cup\{R\}|}$ where $|\{Y\}|$ denotes the size of a set Y.

The *p-value* and the direction of difference from the null hypothesis (that the positions of $\{Q\}$ and $\{R\}$ are independent) are obtained by permutation. Each permutation randomizes the query intervals uniformly across the chromosome, maintaining the spacing between intervals.

Comparisons limited to genomic subsets

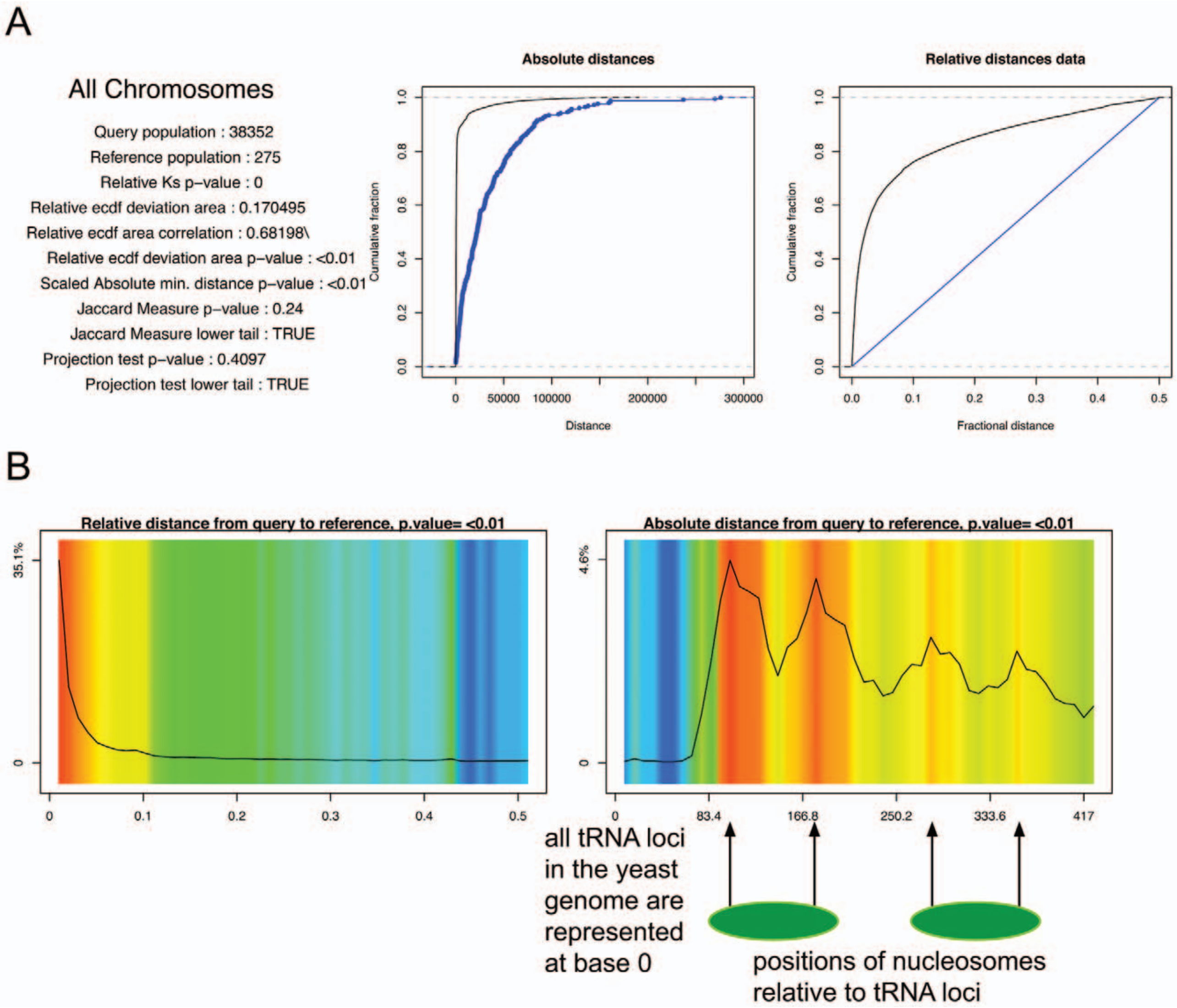
An investigator may want to explore correlations within defined intervals rather than genomewide; for example, when looking at binding sites within and very close to genes, the correlation between these sites will be extremely high genomewide because they are constrained to small intervals, but upon examination of the sites within genes, there may be no correlation at all. For this



```
All Chromosomes

Query population : 29944
Reference population : 60891
Relative Ks p-value : 0
Relative ecdf deviation area : 0.0112594674911978
Relative ecdf area correlation : -0.0443966166449247
Relative ecdf deviation area p-value : <0.01
Scaled Absolute min. distance p-value : <0.01
Jaccard Measure p-value : <0.01
Jaccard Measure lower tail : FALSE
Projection test p-value : 0
Projection test lower tail : FALSE
```

Figure 7. Human genomic CpG islands from Wu et al [9] correlated with the positions of coding sequences in the human genome. Query and reference colors as in Figure 1.
doi:10.1371/journal.pcbi.1002529.g007



```
All Chromosomes

Query population : 38352
Reference population : 275
Relative Ks p-value : 0
Relative ecdf deviation area : 0.170495
Relative ecdf area correlation : 0.68198\
Relative ecdf deviation area p-value : <0.01
Scaled Absolute min. distance p-value : <0.01
Jaccard Measure p-value : 0.24
Jaccard Measure lower tail : TRUE
Projection test p-value : 0.4097
Projection test lower tail : TRUE
```

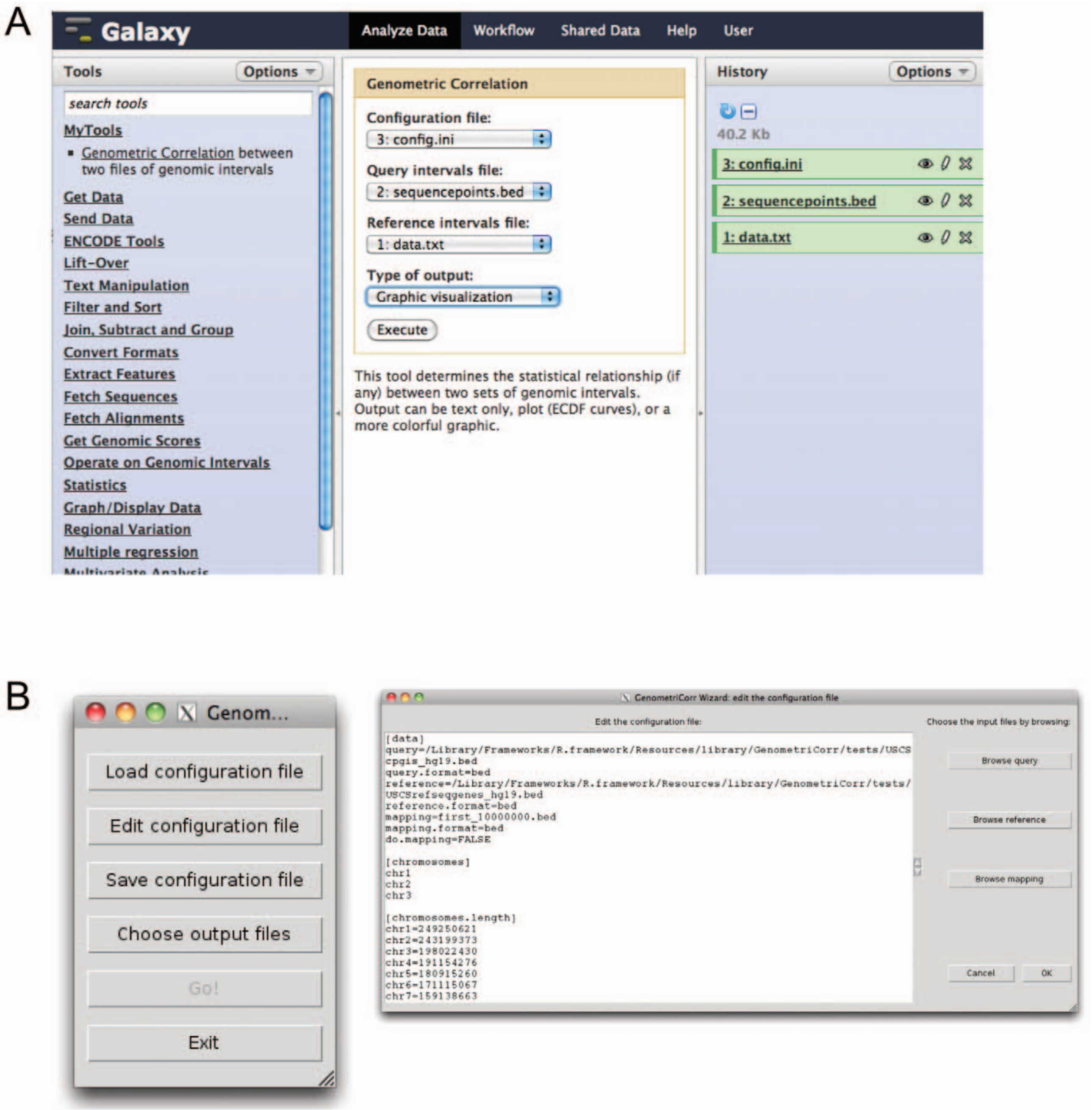
Figure 8. Ty1 retrotransposon insertion sites vs tRNA genes in the yeast genome. (A) ECDF plots (B) Graphic display. Arrows mark Ty1 insertion sites at nucleosome-occupied positions near tRNA genes. Nucleosomes are in green. The colored graph contains several regions of high observed/expected Ty1 insertions (red colors), and the black line indicates a high density of Ty1 insertions, as well, in these regions. Relative to the tRNA position, the Ty1 insertion sites are most dense inside the nucleosome occupied regions. Query and reference colors as in Figure 1.
doi:10.1371/journal.pcbi.1002529.g008

reason we provide methods that restrict the correlations to intervals defined by the investigator and that can be set up from within the configuration file.

Results

We tested the software on our own high-dimensional data, a set of cloned insertion sites of an exogenously supplied Ty1 retrotransposon in the gene-rich yeast genome, for which we were trying to determine Ty1 targeting specificity. After sequencing and mapping it was clear that the insertions cluster near tRNA genes but do not generally insert into these genes, as seen in Figure 8A. Figure 8B displays the very complex relationship between Ty1 and tRNA promoters; the insertions occur at very specific points along nucleosome-bound DNA and follow a reproducible periodic pattern. On further examination we were able to map the insertion sites precisely to the nucleosome surface, as we have previously reported [13].

The examples provided here illustrate the range of biological questions that can be addressed with our software. A particularly compelling feature of the package is that negative correlations (overlap or proximity much less than expected if the query and reference were unrelated) are reported, meaning that correlations between factors that act at a distance are detectable. Also, we



Configuration file: 3: config.ini

Query intervals file: 2: sequencepoints.bed

Reference intervals file: 1: data.txt

Type of output: Graphic visualization

This tool determines the statistical relationship (if any) between two sets of genomic intervals. Output can be text only, plot (ECDF curves), or a more colorful graphic.

```
[data]
query=/Library/Frameworks/R.framework/Resources/library/GenometriCorr/tests/USCS
cpgis_hg19.bed
query.format=bed
reference=/Library/Frameworks/R.framework/Resources/library/GenometriCorr/tests/
USCSrefseqgenes_hg19.bed
reference.format=bed
mapping=first_10000000.bed
mapping.format=bed
do.mapping=FALSE

[chromosomes]
chr1
chr2
chr3

[chromosomes.length]
chr1=249250621
chr2=243199373
chr3=198022430
chr4=191154276
chr5=180915260
chr6=171115067
chr7=159138663
```

Figure 9. A) The Galaxy interface to GenometriCorr. B) The Tk interface to GenometriCorr. Instructions for using both are found on the website.

doi:10.1371/journal.pcbi.1002529.g009

observe that absolute and relative distances are both important and may measure different phenomena; thus the software provides appropriate tests for these correlations as well.

We do not intend for our software to supplant the role of either the biologist or the statistician in a team of investigators working on high throughput sequencing or microarray data. Rather, by determining the statistical significance of genomewide interactions, the software serves as a hypothesis generator, enabling all investigators to begin validating associations that are likely to be real, much earlier than they would have otherwise.

We do not provide a built-in method for retrieving query and/or reference features that may conform to a configuration suggested by the correlation methods. As we provide methods to use the software from within the Galaxy interface (below, and Figure 9), users with minimal computational experience can create any desired subsets using the many tools available in that environment. More computationally experienced users will have no trouble extracting query and reference intervals and overlapping these intervals as suggested by the correlation output.

GenometriCorr can be customized for use with any genome and any type of point or interval data.

Availability and Future Directions

GenometriCorr is available, along with examples and installation guidelines, from http://genometricorr.sourceforge.net/. The software is written in R and can be used from the R command line, through a Tk graphical interface, or through the Galaxy interface; all of these options are documented on the site.

In future work we plan to implement the generalized Jaccard measure, which can handle sparsely distributed query and reference sets. Moreover, the generalized Jaccard measure can account for intervals that have a weight or other numerical value, in addition to coordinates. This weight can denote anything from multiplicity of coverage to experimental confidence.

Author Contributions

Conceived and designed the experiments: AF LM LMC YM AAM VJM SJW. Performed the experiments: AF LM SJW. Analyzed the data: AF LM SJW. Contributed reagents/materials/analysis tools: AF LM LMC SJW. Wrote the paper: LMC SJW.

References

1. Bird AP (1986) CpG-rich islands and the function of DNA methylation. Nature 321: 209–213.
2. Giles KE, Gowher H, Ghirlando R, Jin C, Felsenfeld G (2010) Chromatin boundaries, insulators, and long-range interactions in the nucleus. Cold Spring Harb Symp Quant Biol 75: 79–85.
3. Bickel PJ, Brown JB, Huang H, Li Q (2009) An overview of recent developments in genomics and associated statistical methods. Philos Transact A Math Phys Eng Sci 367: 4313–4337.
4. Bickel PJ, Boley N, Brown JB, Huang H, Zhang NR (2010) Subsampling methods for genomic inference. Ann Appl Stat 4: 1660–1660–1697.
5. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol 5: R80.
6. Gangadharan S, Mularoni L, Fain-Thornton J, Wheelan SJ, Craig NL (2010) Inaugural article: DNA transposon hermes inserts into DNA in nucleosome-free regions in vivo. Proc Natl Acad Sci U S A 107: 21966–21972.
7. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, et al. (2010) Galaxy: A web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol Chapter 19: Unit 19.10.1–21.
8. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61–5.
9. Oeckinghaus A, Ghosh S (2009) The NF-kappaB family of transcription factors and its regulation. Cold Spring Harb Perspect Biol 1: a000034.
10. Zhang Y, Romanish MT, Mager DL (2011) Distributions of transposable elements reveal hazardous zones in mammalian introns. PLoS Comput Biol 7: e1002046.
11. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, et al. (2010) The NIH roadmap epigenomics mapping consortium. Nat Biotechnol 28: 1045–1048.
12. Wu H, Caffo B, Jaffee HA, Irizarry RA, Feinberg AP (2010) Redefining CpG islands using hidden markov models. Biostatistics 11: 499–514.
13. Mularoni L, Zhou Y, Bowen T, Gangadharan S, Wheelan S, et al. (2012) Retrotransposon Ty1 integration targets specifically positioned asymmetric nucleosomal DNA segments in tRNA hotspots. Genome Res 22: 693–703.